# FACE SKETCH RECOGNITION USING DEEP LEARNING

A MAJOR PROJECT-II REPORT
SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF THE DEGREE
OF
MASTER OF TECHNOLOGY
IN
**INFORMATION SYSTEMS**

Submitted by:
**OM AMRIT**
**2K22/ISY/10**

Under the supervision of

**Dr. VARSHA SISAUDIA**

**DEPARTMENT OF INFORMATION AND TECHNOLOGY**
**DELHI TECHNOLOGICAL UNIVERSITY**
**(Formerly Delhi College of Engineering)**
**Bawana Road, Delhi-110042**

**May, 2024**

**DEPARTMENT OF INFORMATION TECHNOLOGY**
DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)
Bawana Road, Delhi-110042

## CANDIDATE'S DECLARATION

I, Om Amrit, 2K22/ISY/10 student of M.Tech (IT), hereby declare that the Major Project-II dissertation titled "**Face sketch Recognition Using Deep Learning**" which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi
Date: 31/05/2024

Om Amrit
2K22/ISY/10

**DEPARTMENT OF INFORMATION TECHNOLOGY**
DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)
Bawana Road, Delhi-110042

# CERTIFICATE

I hereby certify that the Major Project-II dissertation titled “**Face Sketch Recognition Using Deep Learning**” which is submitted by OM AMRIT, 2K22/ISY/10, Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is a record of the project work carried out by the students under the guidance of Dr. Varsha Sisaudia. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this university or elsewhere.

Varsha

Place: Delhi
Date: 31/05/2024

**Dr. Varsha Sisaudia**
SUPERVISOR
**Asst.Professor**
Department of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY

# ABSTRACT

Face Recognition is used to recognise or identify the faces through photos or videos. Today, Face Recognition has a very wide application involving various fields. The accuracy, speed and easy implementation behind the facial recognition system makes it even more useful in day-to-day work. Beyond just unlocking phones and laptops, the face recognition and identification is highly used among security and surveillance.

Recognition of face from sketch is very important in police verification to track any person or criminals. The face is drawn by sketch artist as described by the eyewitness and then the face is recognised through the police database. There are various methods to automatically identify the sketches from a large database has been implemented but using the forensic sketches often reduce the performance.

Many techniques have been used to automatically identify subjects described by eyewitnesses in sketches; however, these techniques frequently perform worse when extended galleries resembling law enforcement mug-shot galleries and real-world forensic sketches are used. Despite deep learning's success in many application areas, including traditional face recognition, not much effort has been done to apply it to face photo-sketch recognition. This is mostly because there aren't enough sketch pictures accessible to train big networks in a reliable manner. Using deep learning techniques, this thesis attempts to address these problems.

# ACKNOWLEDGEMENT

I am very thankful to Dr. Varsha Sisaudia (Assistant Professor, Department of Information Technology) and all the faculty members of the Department of Information Technology at Delhi Technological University. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

OM AMRIT
2K22/ISY/10

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION

Face recognition systems use facial recognition data to attempt to verify an individual using picture or video evidence. Compared to other biometric identification technologies, it has five advantages: non-intrusiveness, convenience, friendliness, non-contact, and scalability. These days, facial recognition is used for criminal identification, security, payments, and entry to buildings. Because criminal suspects may purposefully evade the monitor's range, it is impossible to obtain a clear front-facial image for criminal identification using a typical face recognition system. For instance, to lessen the likelihood of being identified by an electronic system, criminal suspects frequently conceal any distinguishing traits on their faces.

Furthermore, a suspect's photo was taken somewhat later than any other photo in the dataset, making identity theft possible. In the event that a suspect's photo is not available, taking a drawing is a crucial backup plan for bolstering witness recognition. Software can be used to generate the sketch, which is made in accordance with the eyewitness descriptions. From the eyewitnesses' descriptions, it can identify the location and relationships in the evidence and remove irrelevant and misleading elements. Additionally, it might have a tip that would help investigators ask witnesses or suspects and expand their picture of the murder scene. For law enforcement, the automatic facial sketch recognition system is a useful tool.

A face detection system is essentially a computer programme that can recognise or authenticate an individual from a digitized photo or video. It can be compared to other biometrics like fingerprint or eye iris recognition systems and is typically utilised in security systems. In most cases, effective uses of picture understanding and analysis has been facial detection.

There are many instances when all the evidence is missing, but there may be a person who can be eyewitness of the crime. In these situations, a sketch artist is frequently required to collaborate with the eyewitness to sketch the offender as per the details given by the person. When the forensic sketch is complete, it is shared with police or law officials and the media in the hopes that anyone may be familiar with the suspect.

The existing procedure takes a long time. Instead, if there were an automated method for retrieving prospective suspects from police mug shots, it would greatly minimise the amount of time required and simplify the process of identifying suspects.

**Some areas of application of face recognition:**

- For identification of person.
- Healthcare
- Immigration
- In mobile and computers.
- For criminal search

Sketching techniques are used to identify accused from an observer recall. When the sketch is ready then it is matched with the photos available with the department to recognize the suspect.

## 1.1 Face sketch photo recognition

The process of matching a sketched facial image to a set of face images is known as face sketch recognition. Face sketches can be created from descriptions of features or they can be entirely hand-drawn. An artist creates hand-drawn sketches by hand, while software creates composite sketches. Face drawings and images are separated by a significant modality gap that arises from their distinct generating mechanisms, such as a different method of representing the image. The item is projected utilizing the shadow, space, and perspective method to create photos. Sketch uses the lines' sparsity to create the illusion of three dimensions. Nonetheless, due to variations in recollections, certain aspects might be overlooked or overstated.

As a result, there is a difference in texture and shape between the designs and the matching images. The primary obstacle is from the disparity in feature representation between sketches and images, which can be attributed to the modality gap and the heightened description of the sketch. Transferring the facial shot and the sketch to the same medium was the early recognition approach used to minimize the discrepancy. The matching facial photo and the current facial drawing serve as the method's input.

The machine learning approach can then be used to learn the relationship between the two modalities that can be utilized to synthesize a pseudo image.

The characters in the photographs from the collection are used to create the synthesized pseudo image. Since the created image does not exist in the dataset, it is referred to as a pseudo image. The ability to distinguish between a photo and a drawing is superior than that of a pseudo photo and a sketch because some facial features are acquired during the creation of the pseudo image.

### 1.1.1 Types of Sketches

There are three categories of facial sketches, as follows[1]:

1. **Viewed Sketches:** The sketches are created by viewing the image directly. Since they are not dependent on the description, they are not important to forensic. The use of such sketches in law enforcement situations is therefore prohibited.
2. **Forensic Sketches:** All those are depictions of the face created by forensic artist using a witness' descriptions as a guide. forensics drawings have been employed in police prosecutions for a very long time.
3. **Composite Sketches:** In this, using the software the sketches of the faces are created and can be allowed to select component for faces. Such kinds of sketches have proven to be a well-liked and less expensive substitute for forensic sketching.

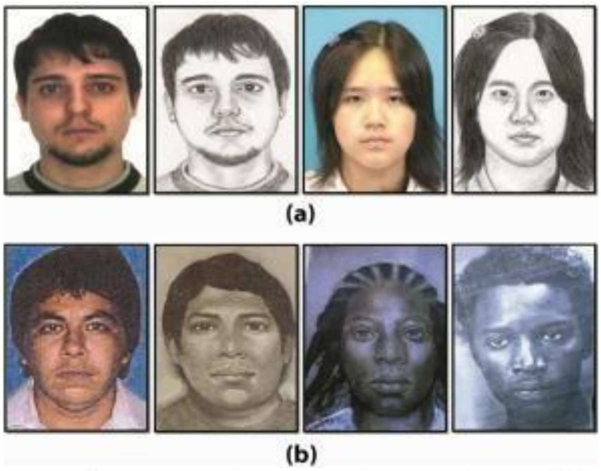

**Fig 1.1.1:** Face photo example

### 1.1.2 Difficulties in face sketch recognition:

Essentially, there are two distinct modalities that include images and sketching. The task of matching a picture with a sketch or matching a sketch with picture is thus challenging and complex. The main challenges encountered when comparing face sketches to pictures are[2]:

- Face sketch recognition is far more difficult than typical recognition of faces from photo images because of the significant drawings' variations from photographs and the cognitive processes involved in drawing sketches, both of which are unknown.
- Comparing the texture of the patches drawn in pencil on paper to that of a photograph of a person, the patches are different. Artists frequently add some shadow texture to sketches to represent the 3D shading information.
- A sketch contains shape warping since it emphasises some distinguishing facial traits like a cartoon.

Various methods have been used to automatically detect the subjects in sketches that matches whatever witness have described. Although its popularity in many application like face recognition systems, little effort has been done to apply deep learning in recognition of face from sketches.

Mostly the causes is lack of publically available photo-sketch pairs, which makes it difficult to train deep networks reliably and prevent problems like over-fitting and local minima . Additionally, a deep network finds it challenging to learn features because there is often only one sketch per topic.

## 1.2 Research Aims and Hypothesis

The recognition of facial sketches has three obstacles. The first difficulty is the discrepancy in the features that are represented between a facial shot and a sketch, which may be observed by comparing their modalities. All types of sketches reflect the subjective opinion and painting style of the creator. Compared to other image sketches, the facial sketch has a more intricate structure. Certain facial features, such the human face's sides and front, its prominent nose, it eye sockets, and cheekbones, are challenging to depict in two dimensions on two-dimensional drawings.

For this reason, painters use chiaroscuro and structural sketching to depict three-dimensional effects. To specify geometric representations in stereo, a structural sketch is employed. In contrast, chiaroscuro refers to the use of bright and dark lines made of the colors black, white, and grey as a crucial technique to accentuate the depth of each face feature. Otherwise, the artist typically uses perspective sketching to depict the shape's structure. For instance, in the front facial sketch image, the artist removes one whole nostril to depict the stereo of various nose kinds. Therefore, in order to emphasize the three dimensions of the facial sketch, delicate elements will be overlooked.

Even so, the templates for each facial characteristic in the component sketches produced by software retain subtle details to depict stereo.

The second problem is that training deep learning models requires really big datasets, even bigger than any photo-sketch dataset we've seen before. Deep learning improves recognition accuracy by figuring out abstract details using a complex system that works in a different way from traditional methods.

Small dataset network performance may be hampered by overfitting since the amount of parameters is excessive for training. Directly updating the model for every neuron's weight when the training set is sparse typically results in overfitting and lowers network performance. Additionally, every node has some hidden features. Deep learning systems act like "black boxes," making it challenging to gather historical data for network training.

The third issue is that many facial sketch datasets only have one photo and one sketch for each person. This makes it tough for deep learning models to give accurate results across different groups of people. It's hard to make powerful, widely useful models using regular deep learning methods.

**Hypothesis 1 (Extract effective features):** For all facial sketch datasets, traditional face photo-sketch identification cannot yield a high recognition rate due to the inability of derived features to mitigate the impact of photos with varying modalities. In one-shot recognition[4], for any classification test where we have one example of each class, the Siamese neural network performs well; nevertheless, in encoding the features of facial photos and sketches, it is more complex than other images.

We conjecture that the autoencoder network functions [5] as a channel structure for the purpose of extracting features that enable a more effective comparison of the distance between facial sketches and images using the contrastive loss .

**Hypothesis 2 (More attention to the same of regions):** The Siamese neural network consist two same type of samples are represented in the same part of the embedding space by using the same process for every branch.

It facilitates the use of the network of shared parameters for facial sketch dataset recognition. But there are differences between how characteristics are depicted in facial sketches and images. Not every related characteristic in a facial drawing and photo can be learned by the Siamese neural network.

Our theory is that in order to distinguish distinct people in photos and sketches, the attention mechanism[6] looks for similar portions of the image that contain a wealth of information in an effort to build and enhance recognition accuracy. A spatial pyramid pooling layer is applied after the image has been cropped in order to lessen the amount of information lost during the pre-processing stage.

# CHAPTER 2: LITERATURE REVIEW

## 2.1. Face recognition Technology

Since the beginning of time, people have been sketching one other's features in an attempt to freeze a moment in time, whether it be an image or a person. Rather of aiming for an exact likeness, the image searches for the presence of a person . Although no two looks are exactly same, a system of classification can be used to identify a single person from thousands of photographs based on their facial features and head shapes. These recognizable characteristics allow us to cognitively encode and store facial images for subsequent retrieval. When there is little evidence and the identity of the criminal is unclear, law enforcement organizations employ sketches to help them with their investigations.

Known as the "Portrait Parle" or "speaking likeness," Alphonse Bertillon, who is frequently referred to as the pioneer of scientific detection, created an identifying technique in the 1880s. This method comprised a collection of facial traits extracted from photos along with accompanying descriptive text. Bertillon had originally intended the catalogue to serve as a means of identification to aid in the identification of local convicts, but it was also discovered to be helpful in gathering descriptions of suspects who were not well-known. Bertillon's classification served as the foundation for contemporary recall systems, which aided artists in creating sketches, composite kits, and computer systems.

Following a bombing in a Wall Street office in 1920, a preliminary composite sketch was created. A witness from a neighboring blacksmith's forge who had shod a stranger's horse and saw it with something covered in the rear of his wagon was located during the investigation. The blacksmith claimed in an interview that he could have given an artist enough information about the stranger's face to enable them to sketch a portrait of him. A commercial artist was commissioned to create a sketch that bore such a striking resemblance to the stranger that he was subsequently apprehended.

## 2.2 Literature review on face photo sketch recognition

Different generating techniques are used to produce photos and sketches, which represent distinct types of representation. This isn't a result of the artist's sketching abilities; rather, it's because no artist obtains sufficient reliable information from the victims or witnesses. Because extreme occurrences can cause great trauma, witnesses or victims may forget all or part of the encounter; hence, people "seal up" the memory of the event as a self-protective technique. The difference between a photo of a face and its corresponding sketch presents the biggest obstacle to facial sketch recognition. Deep learning techniques or conventional methods are used for facial sketch recognition. The conventional approaches can be further separated into feature-based, common space-based, and synthesis-based approaches.

### 2.2.1 Methods based on Synthesis based

One useful strategy for lessening the modality gap between face sketches and pictures is the synthesis-based method. By producing a high-quality pseudo image, this approach successfully closes the modality gap. Tang and Wang[8] proposed the first automatic facial sketch image retrieval technique. This involves analyzing the covariance matrix of all the faces and using a mathematical transformation called the Karhunen-Loeve Transform to find a set of important patterns known as eigenvectors. This procedure has the benefit of reducing the image's dimensions as well as the quantity of data. The image is then projected onto eigenface space using these eigenvectors.

During the recognition phase, the accuracy of the recognition is calculated by applying the sketch eigenspace, and photo eigenspace sequentially. The accuracy rate of this approach is 73%. After that, they identify using rank-10 accuracy. When a correct sketch appears in the top 10 predictions, it has a rank of 10. On the CUFS dataset, the rank-10 accuracy rises to 96%, surpassing the recognition outcome on the rank-1 accuracy.

However, because principal component analysis (PCA) **[9]** cannot synthesize all of the details of the sketches, particularly when the subject's hair is included, the performance of recognition may be impacted.

Using the idea that each data point and its neighbors are close together on a curved surface (manifold), a method that preserves local relationships calculates a mapping from the original high-dimensional data to a lower-dimensional space. This method considers nearby data points when deciding how to map each point. Kernel-based nonlinear discriminant analysis[20] (KNDA) outperforms PCA and linear discriminant analysis (LDA) in classification tasks. KNDA combines the strengths of LDA with a nonlinear kernel technique.

On the CUFSF dataset, KNDA achieves the highest recognition rate at 87.67%, while LDA achieves 85%, and PCA achieves 64.33%.

### 2.2.2 Methods based on Common space

Projecting face sketches and images into a single feature area is the goal of the common space method. In cross-domain image classification problems that are treated using common space based algorithms. Sharma [11] suggests obtaining linear projections for sketches and pictures of faces by employing Partial Least Squares (PLS). These projections from various modalities are then mapped onto a shared area. For multi-modal recognition, the closest neighbor approach is employed after maximizing the common covariance.

Using a holistic methodology, a rate of 93.6% accuracy is obtained with the CUHK database. In order to create a discriminant common space use the Multi-View Discriminant Analysis (MvDA) approach [12]. This method involves learning different linear transformations for different views simultaneously. It's more effective and versatile because it optimizes a common space that discriminates across multiple views. To improve recognition, it maximizes the distance between images from different modalities using the generalized Rayleigh quotient. This method is tested on three types of facial recognition datasets: the Multi-PIE dataset, the CUHK Facial Sketch FERET dataset, and the Heterogeneous Facial Biometrics dataset.

The best performance is guaranteed by the outcomes of these three datasets. In particular, the rank-1 recognition for photo-sketch and sketch-photo recognition is 53.4% and 55.5%, respectively, using the CUFSF dataset. Nonetheless, sketches share a lot of characteristics, therefore the classifiers might not be very useful.

### 2.2.3 Methods based on feature

In 2010, Klare and Jain[23] proposed the first feature-based technique. This approach extracts gradient information by employing multi-scale LBP [13] and SIFT [14]. Here, LFDA increases the accuracy of recognition. The modality gap between sketches and pictures is not overcome by LFDA, despite the remarkable accuracy attained with this method. This approach tests the validity using an already-existing mug-shot collection. The range of the success rate is 10 to 50. Using a race/gender filter improves accuracy by 18.37% to 44.90%. Compared to Face-VACS [15] and the suggested method without a race/gender filter, this result is superior.

In face photo-sketch recognition, MLBP and SIFT methods struggle when there's a difference between the two modalities. Zhang et al.[16] introduced information theoretic methods to enhance the shared information between facial sketches and photos in the feature space. This descriptor is designed to collect more distinct and useful information to improve recognition accuracy.

### 2.2.4 Method based on Deep-Learning

Deep learning methodologies leverage the hidden layer activations within a neural network to extract low-level, mid-level, and high-level features from the input data. These feature representations play a crucial role in accomplishing the end-to-end tasks for many applications. This can include a holistic facial analysis system that involves tasks such as facial detection, alignment, expression recognition and Convolutional Neural Network (CNN) based model classification. is suggested by DeepFace[17].

Feature selection is carried out by an eight-layer CNN model. In order to increase the robustness of the model, a max-pooling layer is inserted between every two convolution layers. Basic features such as edges and textures are the focus of the convolution layers in the initial three layers. The remaining layers consist of a SoftMax layer, one full connect layer, and three local convolution layers.

The unshared convolution kernel used by local convolution layers gives this method the advantage of requiring fewer parameters. Moreover, the DeepFace[17] method minimizes the number of max-pooling layers to keep texture characteristics from disappearing.

The LFW dataset[21] is used to train and assess the algorithm. With a single CNN model, the recognition accuracy is up to 97% on the front end. To collect advanced feature representation for categorization, the DeepID [18] leverages deep learning. This model uses four convolution blocks: a fully connected layer, a max pooling layer, a convolution layer, and a SoftMax layer acting as a classifier.

Despite sharing structural similarities with the original convolutional neural network, the DeepID model's final convolution layer is utilized to extract complementary and comprehensive feature representations following the input of various facial photos into the CNN model. With the DeepID models, recognition accuracy can reach 97.25% on the LFW dataset, owing to their excellent capacity to generalize. Furthermore, a deep neural network with several hidden layers may learn characteristics quite well.

For a facial photo-sketch recognition task, deep learning techniques can be split into two categories, one category uses synthesis of the pseudo-image.

High-quality photos based on "Perceptual Losses for Real-Time Style Transfer and Super-Resolution" are synthesized by Güçlütürk[19] using a DNN model.Three loss functions are used to create a synthetic image that closely resembles the original: a typical Euclidean loss function, a Euclidean loss function, and the loss function of a single pixel.

The integrated loss function measures the difference between an actual image and one that is anticipated by comparing features and calculating the image's features in pixels using a conventional Euclidean loss function. Finally, the accuracy of the recognition between synthetic color images and real photos is checked using RS-LDA[8]. When color images are synthesized using line drawings, a 99.79% accuracy rate is attained. One drawback, though, is that the model can only produce a single image from a single sketch. Another is that not all photos have precisely synthesized color.

In convolution sketch inversion, the edges of pseudo drawings appear blurry. To create a more realistic image, a regular feed-forward network is employed. This network is trained to understand how to turn a sketch into a colored picture.

.Bromley et al[22]. introduced the first Siamese model in 1994.The CNN model performs well in facial recognition tests, however it is unsuitable for real-time systems or realistic settings.

# CHAPTER 3: METHODOLOGY

## 3.1 Database Used

**1. Person Face Sketches(Kaggle):**Dataset with 21K+ images pair of person face photos and sketches.

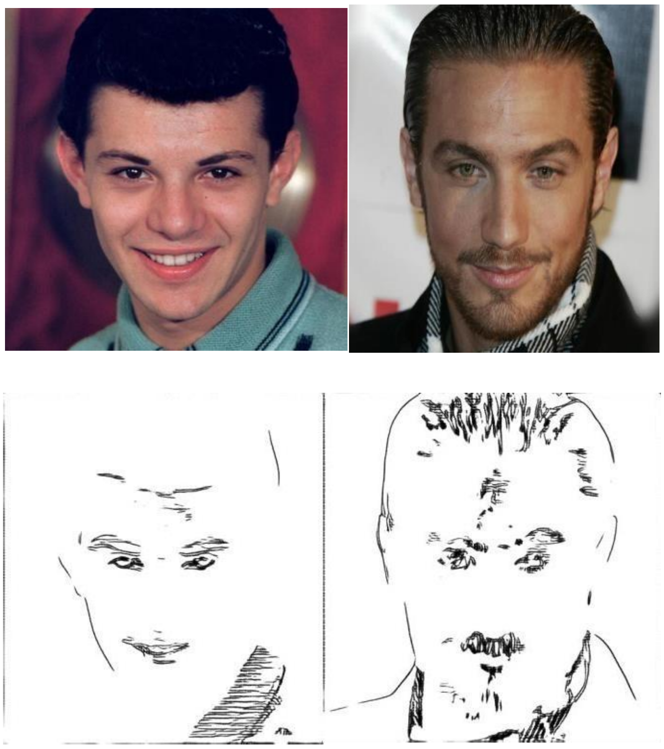

Fig 3.1.1:Person face and sketches(Kaggle)

**2. CUHK Database of face sketch(CUFS):**

- This database is mainly used for research on face sketch synthesis and face sketch recognition.
- It has total of 188 pairs of faces and their sketch which is provided by the Chinese University of Hong Kong (CUHK) student database.

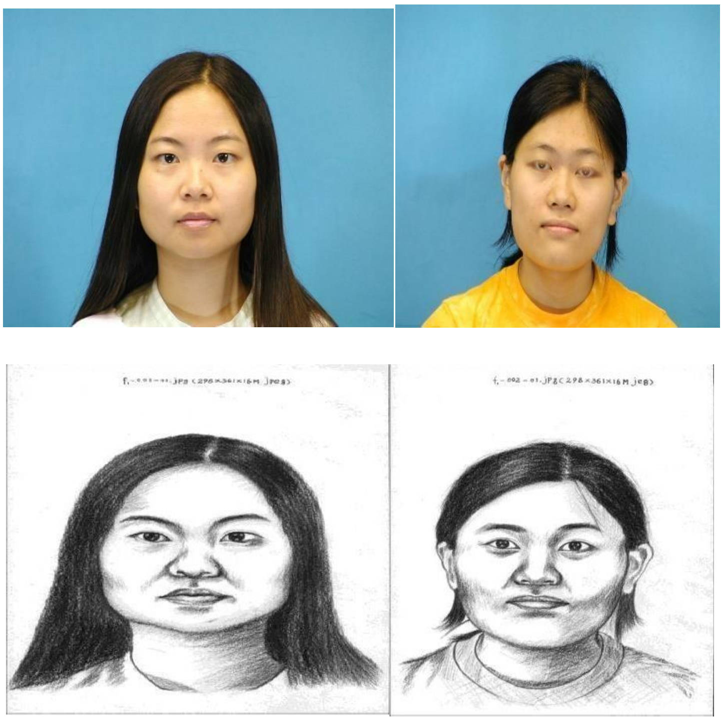
f,-001-01.jpg (298×361×16M jpeg)
f,-002-01.jpg (298×361×16M jeg)

Fig 3.1.2:CUHK face and sketch

## 3.2 An Improved Siamese Network

### 3.2.1 Introduction

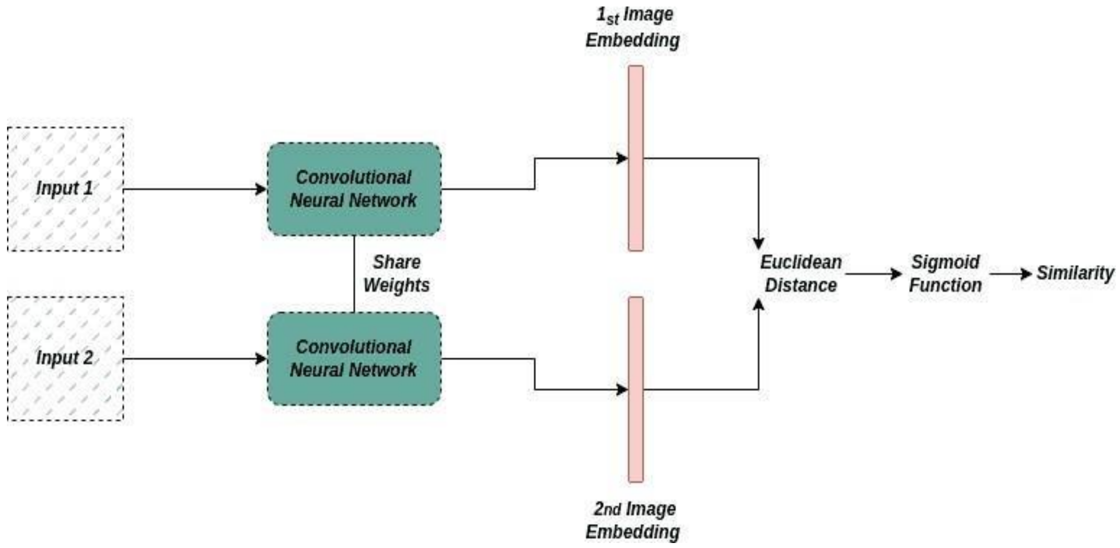

Fig 3.2.1: Siamese Network Architecture

Siamese neural networks are a class of neural network architectures designed to compare and measure similarity between pairs of input samples. The term “Siamese” comes from the idea that the network architecture consists of twin neural networks, which are identical in structure and share the same set of weights. Each network processes one input sample from the pair, and their outputs are compared to determine the similarity or dissimilarity between the two inputs.

The main motivation behind Siamese networks is to learn a meaningful representation of input samples that can capture their essential features for similarity comparison. These networks excel in tasks where direct training with labeled examples is limited or difficult, as they can learn to differentiate between similar and dissimilar instances without requiring explicit class labels.

A Siamese network usually has three key components – shared network, similarity metric, and contrastive loss function in its architecture.

1. **Shared Network**: The Siamese architecture has a central component which is the shared network. This network is charged with extracting significant feature representations from input samples. The shared network is composed of layers made of neural units like convolutional layers, or fully connected layers, which in turn process input data creating fixed length embedding vectors. When the same weights are shared between the two networks, the model learns to extract alike characteristics according to analogous input signals and enabling effective comparison.

2. **Similarity Metric**: After the inputs are processed by the shared network, a similarity metric is used to compare the generated embeddings and measure how similar or different the two inputs are. The choice of similarity metric depends on the specific task and the type of input data. Common metrics include Euclidean distance, cosine similarity,and correlation coefficient. These metrics calculate the distance or correlation between the embeddings, giving a measure of how similar the input samples are to each other.

3. **Contrastive Loss Function**: During training, we utilize a contrastive loss function. This function encourages the network to generate similar representations for similar inputs and different representations for dissimilar inputs. It penalizes the model when the distance or dissimilarity between similar pairs exceeds a certain threshold, or when the distance between dissimilar pairs falls below another threshold. The exact formulation of the contrastive loss function depends on the chosen similarity metric and the desired margin between similar and dissimilar pairs.

During training, the Siamese network learns to optimize its parameters to minimize the contrastive loss and produce discriminative embeddings that effectively capture the similarity structure of the input data.

### 3.2.2 The proposed Siamese network architecture

The suggested Siamese network design uses two comparable convolutional networks as channels to extract information from both the images and sketches of the face. Various photos are fed into these networks, and their weights are shared. By implementing the sharing weights model across the convolutional networks of the two channels, we use the contrastive loss function to translate the features that were extracted individually from each of the two input images into a common space. After each channel's final layer, the outputs of the sub-networks in the Siamese network are combined and trained using the contrastive loss function. The goal is to widen the difference between pairs of images that don't match (negative pairs) and narrow the gap between pairs of images that do match (positive pairs).

It is challenging to extract sufficient textural features for recognition using a convolution network with a small kernel size since the sketch uses monochromatic lines to portray the object's structure.

For instance, the photo image's small circular patches can be identified as human eyeballs based on the surrounding space and the grey-level distribution of the pixels. But because the sketched depiction emphasizes shape, using the retrieved texture data can lead to them not being identified as eyes.

For recognition, we therefore employ a large kernel size to capture more structural characteristics than texture features. The "rectified linear unit" (RELU) activate function is used in all convolution layers to create nonlinear mapping, which maintains a faster learning rate and strengthens representational ability more than other activation functions.

Furthermore, every photo-sketch dataset pertaining to faces that we utilized originated from full-face photos.

As a result, the images don't require any padding to alter the quantity of pixels in the photos and matching designs for the three convolution layers that house the character mappings. This indicates that all dimensions are valid in order for the filter and stride to cover the input image and for the filter window to remain in a valid location inside the input map.

Padding is used to increase the number of pixels for the input pictures in the final four convolution layers. This conveys less information in face images and sketches, even though it might lose certain features on the edge of the picture.

Padding has two benefits: it reduces the number of parameters and maintains enough features for a deeper layer.

### 3.2.3 Loss Function

Loss functions are used in machine learning or deep learning models to measure how much the predicted value differs from the actual value. When the loss function decreases, it means the model is performing better. In our research, we're learning a mapping that aligns various features from different modalities into a shared space. To do this, we calculate the loss function, which measures the distance between a facial photo and a sketch. The objective is to use a constraint condition to distinguish the intra-modal and inter-modal samples. The network is trained using two different loss functions, such as the contrastive, Cross-entropy loss functions.

#### 3.2.3.1 Contrastive loss function

Contrastive loss is a type of loss function often used in Siamese networks to learn how similar or different pairs of input samples are. It helps train the network so that similar inputs end up with embeddings that are close to each other, while different inputs have embeddings that are farther apart. By reducing the contrastive loss, the network learns to create embeddings that accurately reflect the similarities and differences in the input data.

The contrastive loss key components and steps:

1. **Input Pairs**: The contrastive loss function works with pairs of input samples. Each pair includes a similar (positive) example and a different

(negative) example. These pairs are usually created during training, with positive pairs showing similar cases and negative pairs showing different cases.

2. **Embeddings**: The Siamese network processes each input sample using the same network, creating embedding vectors for both samples in the pair. These embeddings are fixed-length representations that capture the important features of the input samples.

3. **Distance Metric**: A distance metric, like Euclidean distance or cosine similarity, measures how different or similar the generated embeddings are. The choice of metric depends on the type of input data and the specific needs of the task.

4. **Contrastive Loss Calculation**: The contrastive loss function calculates the loss for each pair of embeddings. It encourages similar pairs to be closer together and dissimilar pairs to be farther apart.

The contrastive loss function compares how similar a pair of images are. It's defined as follows:

$$L = Y * D^2 + (1 - Y) * \max\,(margin - D, 0)^2$$

Where:

- L: Contrastive loss for the pair.
- D: Distance or dissimilarity between the embeddings.
- y: Label indicating whether the pair is similar (0 for similar, 1 for dissimilar).
- m: Margin parameter that defines the threshold for dissimilarity.

The term $(1 — y) * D^2$ in the loss function penalizes similar pairs if their distance is too large, encouraging the network to bring them closer. The term y * max(0, m

— D)$^2$ penalizes dissimilar pairs if their distance is too small, pushing the network to separate them more.

5. **Aggregating the Loss**: To get the total contrastive loss for the whole batch of input pairs, the individual losses are usually averaged or added up. The choice of method depends on the specific training goal and optimization strategy.

By minimizing the contrastive loss using methods like backpropagation and stochastic gradient descent, the Siamese network learns to create embeddings that clearly show how similar or different the input data is.

The contrastive loss function is key in training Siamese networks, helping them learn useful representations for tasks like image similarity, face verification, and text similarity. The exact setup and parameters of the contrastive loss function can be adjusted based on the data and the task's needs.

#### 3.2.3.2 Cross Entropy Loss Function

The difference in cross-entropy between two random variables is called a loss. It displays the outcomes by measuring the variables to extract the differences in the information they carry.

However, let's first talk over loss functions in brief before getting more specific.

According to their outputs, we divide them into two categories:

- Loss functions for classification
- Loss functions for regression

Classification and regression are the two categories that comprise the supervised learning principle. Additionally, the results of classification tasks are categories, such as dogs and cats, but the results of regression tasks, for instance, are numbers.

Cross-entropy loss is the most widely used loss function in classification tasks since it allows us to define the difference between the predicted probability and our desired outcome and determine how accurate our machine learning or deep learning model is.

Your model's performance is measured by the cross-entropy loss function, which converts its variables into real values and calculates the "loss" that results from doing so. The greater the disparity between the two, the greater the amount of loss.

Cross-Entropy Loss Function Types:

In machine learning and deep learning classification problems, there are two main types of the cross-entropy loss function that we identify, which are as follows:

- Binary cross entropy
- Categorical cross entropy

**Binary Cross Entropy**

Binary cross-entropy (BCE) is a loss function often used in binary classification tasks. It evaluates how well a classification model works when its output is a probability ranging from 0 to 1. BCE computes the loss by comparing the predicted probability with the true label, which can be either 0 or 1.

The formula for binary cross-entropy is:

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i \log (p_i) + (1 - y_i )\log (1 - p_i))$$

Where:

- N is the number of samples.
- $y_i$ is the actual label for the $i$-th sample (0 or 1).
- $p_i$ is the predicted probability for the $i$-th sample (ranging from 0 to 1).

**Categorical Cross Entropy**

Categorical cross-entropy, also called softmax loss or multi-class cross-entropy, is a loss function for multi-class classification tasks. It evaluates how well a classification model performs when its output is a probability distribution across several classes. The goal is to compare the predicted probability distribution with the actual distribution (one-hot encoded labels).

The formula for categorical cross-entropy is:

$$L = \frac{1}{N} \sum_{i=1}^{N} \square \sum_{j=1}^{C} = y_{ij} \log (p_{ij})$$

### 3.2.4 Model Implementation And Result

To implement a face-sketch recognition system using an improved Siamese network, we will follow these detailed steps:

**Data Preparation:**

- Organize the data into train, validate, and test directories, each containing face and sketch subdirectories.
- Ensure each image in the face directory has a corresponding sketch image in the sketch directory.

**Data Augmentation and Preprocessing:**

- Create a data generator to handle the loading and preprocessing of images in batches to avoid memory issues.

**Model Architecture:**

- Build a base convolutional neural network to extract features from images.
- Create a Siamese network using the base model to compute the similarity between face and sketch images.

**Loss Functions:**

- Define contrastive, cross-entropy, and hinge loss functions for training.

**Custom Metrics:**

- Define a custom accuracy metric to evaluate the model performance.

**Training the Model:**

- Compile and train the model with each of the loss functions using the training and validation data.

**Evaluating the Model:**

- Evaluate the trained models on the test data and compare their performance.

**Result:**

Model is trained on different loss functions with 20 epochs:

| Dataset | Cross entropy loss | Contrastive loss |
|---|---|---|
| CUHK | 56.35 | 59.28 |
| Person face(Kaggle) | 73.28 | 86 |

Table 3.2.1 Accuracy Result on Siamese Network

**Result on CUHK Dataset:**

**1) Using Entropy Loss**

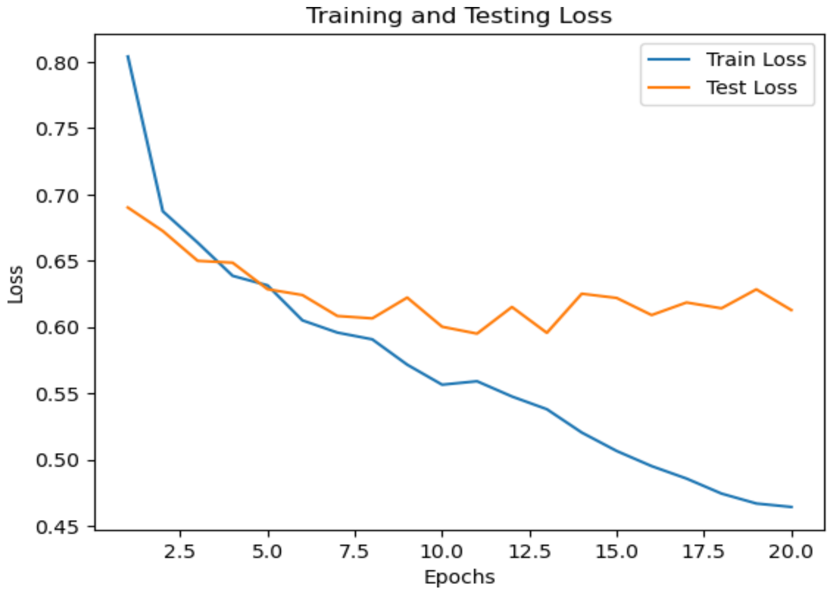

Fig 3.2.2:Training and Testing Loss on CUHK dataset using entropy loss

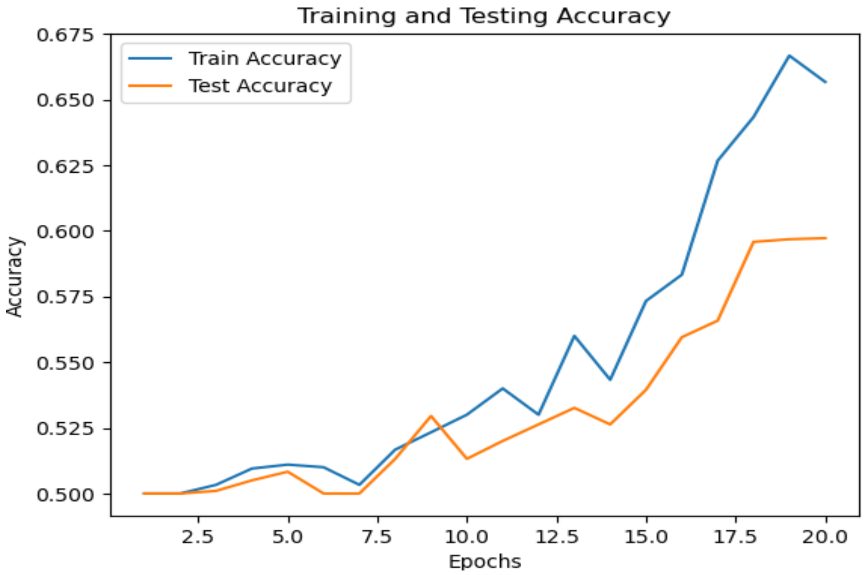

Fig 3.2.3: Training and Testing Accuracy on CUHK dataset using entropy loss.

### 2) Using Contrastive Loss:

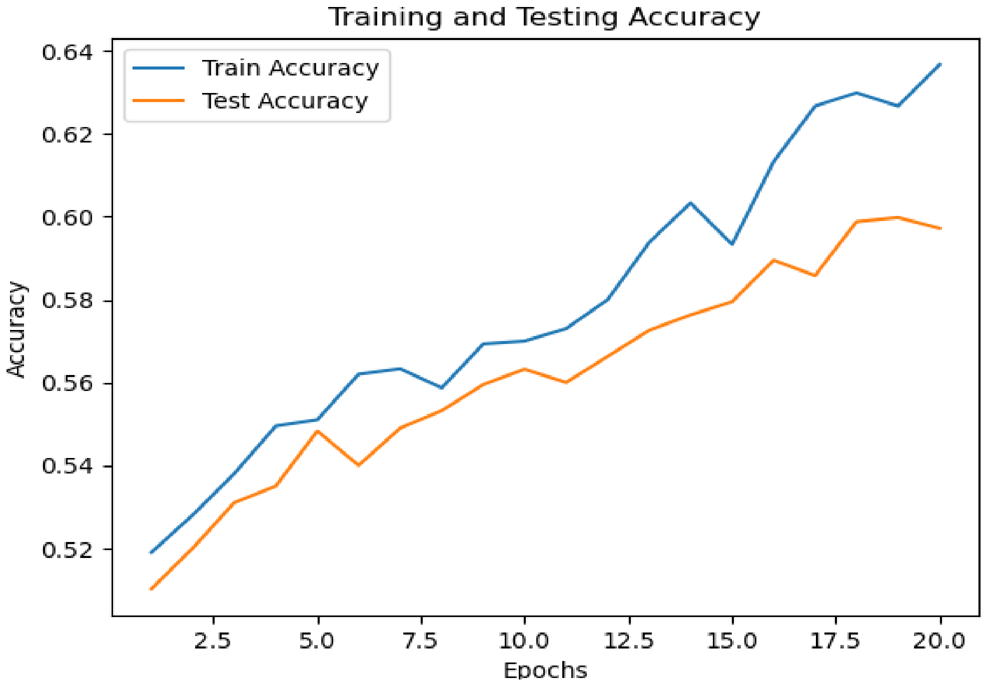

Fig 3.2.4: Training and Testing Accuracy on CUHK dataset using contrastive Loss

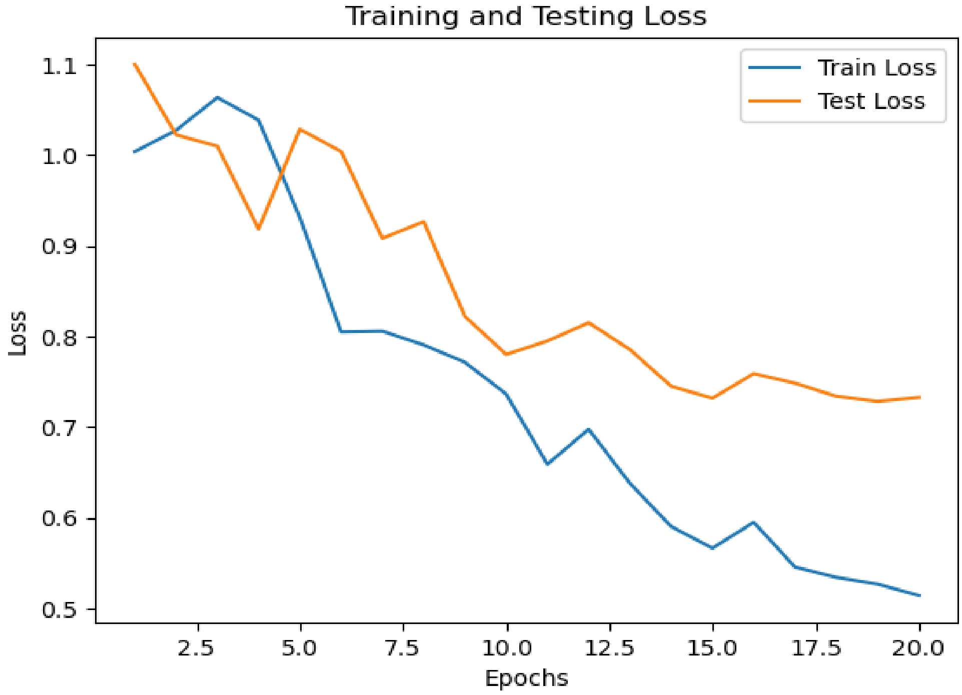

Fig 3.2.5: Training and Testing Loss on CUHK dataset using contrastive Loss.

**Result on Face Sketch Dataset:**

1) **Using Contrastive Loss**

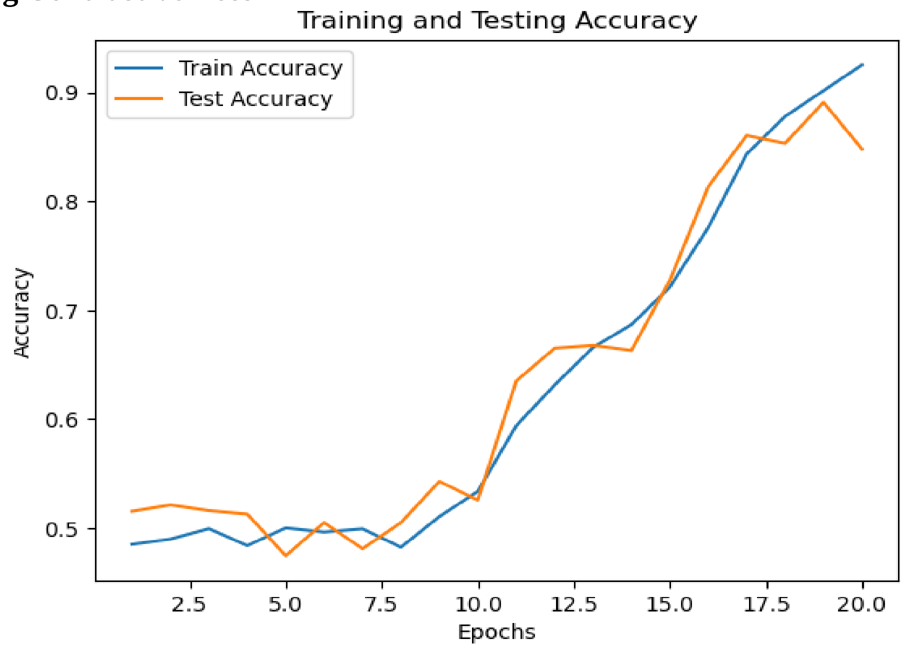

Fig 3.2.6: Training and Testing Accuracy on Face Sketch dataset using entropy Loss.

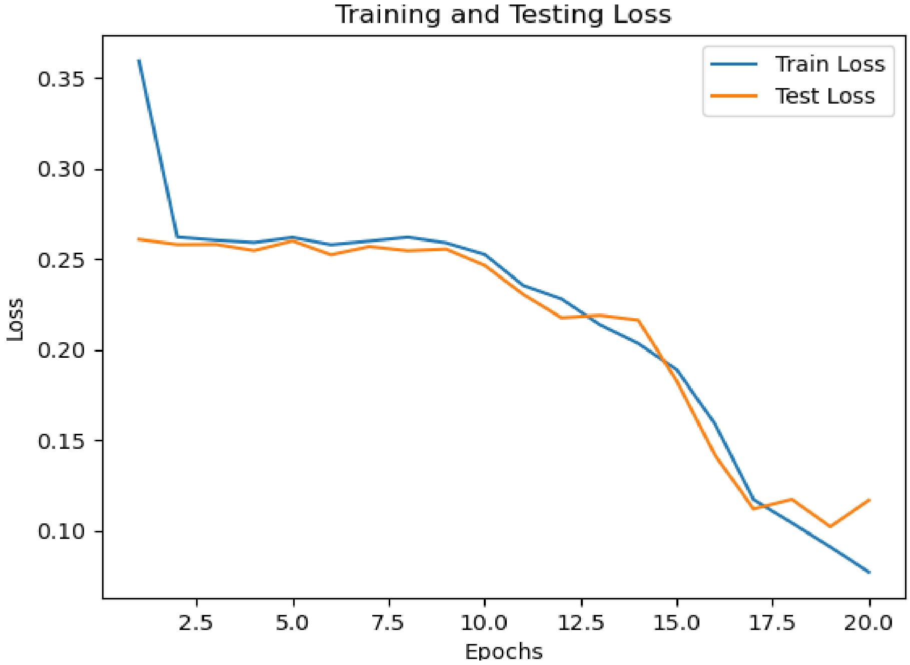

Fig 3.2.7:Training and Testing Loss on Face Sketch dataset using entropy Loss.

### 2) Using Entropy Loss

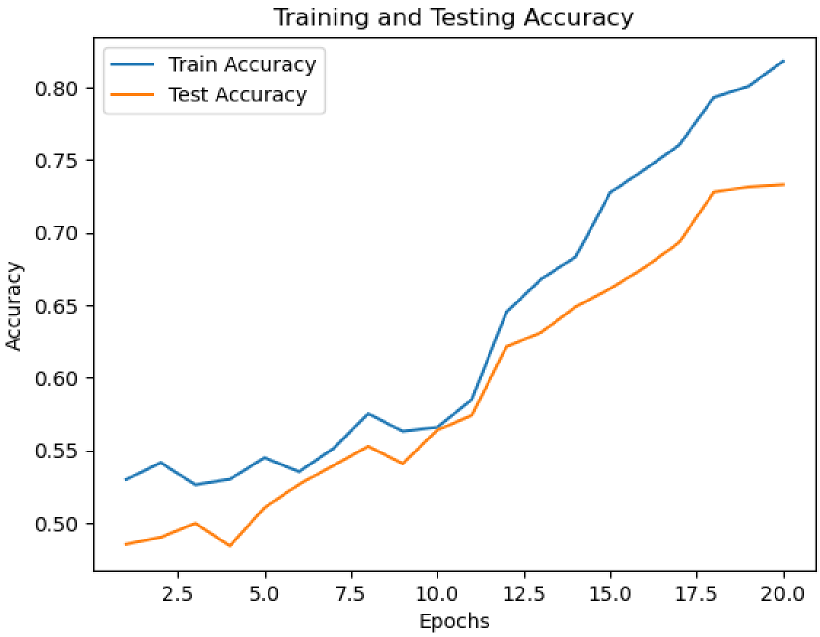

Fig 3.2.8:Training and Testing Accuracy on Face Sketch dataset using entropy Loss.

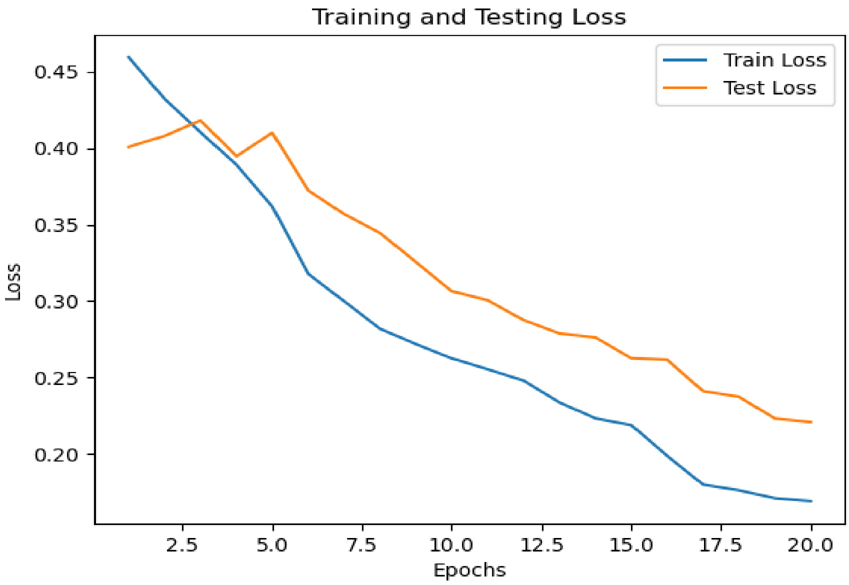

Fig 3.2.9 : Training and Testing Loss on Face Sketch dataset using entropy Loss.

## 3.3 Attention Triplet Network for Face Sketch Recognition

### 3.3.1 Triplet Network Introduction

A Triplet Network is a neural network that learns to recognize similarities and differences between images by comparing triplets of images (anchor, positive, negative). It uses a shared network to create embeddings and a triplet loss function to train itself to keep similar images close and dissimilar images far apart. This makes it very useful for tasks like face recognition and image retrieval.

**Key Components of a Triplet Network**

1. **Anchor, Positive, and Negative Samples:**
   - **Anchor:** The reference examples.
   - **Positive:** An example similar to the anchor.
   - **Negative:** An example dissimilar to the anchor.

2. **Embedding Network:**

   A shared neural network that maps the anchor, positive, and negative samples to a feature space.

3. **Triplet Loss Function:**

   The goal of triplet loss is to create a space where similar samples are closer together and different samples are further apart. Triplet loss models do this by embedding samples so that those with the same label are closer than those with differentlabels.

   Because of this, the triplet loss architecture uses the ideas of similarity and dissimilarity to teach us distributed embedding.

The loss function promotes a certain margin of separation between the anchor and the positive being less than the distance between the anchor and the negative:

$$L = \max(0, d(a,p) - d(a,n) + \alpha)$$

Where:

here $d(a,p)$ and $d(a,n)$ are the distances between the anchor-positive and anchor-negative pairs, respectively, and $\alpha$ is a margin.

**Attention Mechanism:**

An attention mechanism helps the network focus on the most relevant parts of the input when learning the feature embeddings. This is particularly useful in scenarios where parts of the input may carry more important information than others .

### 3.3.2 Attention Triplet Network Architecture:

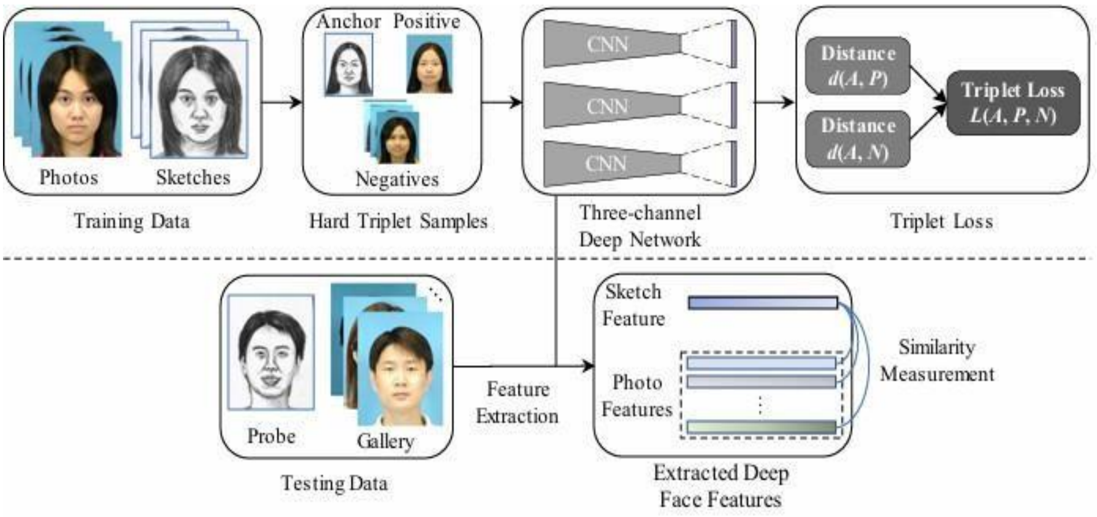

Fig 3.3.1:Diagram of face sketch recognition using Triplet Network.

#### 1. Attention Module:

The attention module is typically placed before or within the embedding network. This module learns to weigh different parts of the input data differently based on their importance. The module can be implemented using various attention mechanisms, such as:

a) Self-Attention:- Computes attention weights based on the input itself.
b) Cross-Attention:- Computes attention weights based on a different input.

#### 2. Embedding Network with Attention

The embedding network processes the input data, now enhanced with attention weights, to produce more discriminative feature embeddings. This network may consist of convolutional layers, recurrent layers, or other neural network architectures suitable for the task.

#### 3. Triplet Loss With Attention:

The triplet loss function remains the same but is applied to the embeddings produced by the attention-enhanced embedding network.

### 3.3.3 Model Implementation and Result

**Working of An Attention Triplet Network:**

1. **Input Processing:**

   The anchor, positive, and negative samples are fed into the attention module. The attention module assigns weights to different parts of the input data, emphasizing the most important features.
2. **Feature Embedding:**

   The weighted inputs are then passed through the embedding network. This network generates feature vectors for the anchor, positive, and negative samples. These feature vectors represent the important characteristics of each sample.
3. **Loss Calculation**

   The triplet loss is computed based on the distances between the feature vectors of the anchor-positive and anchor-negative pairs.
4. **Training**:

   The network is trained to minimize the triplet loss, thereby learning a feature space where similar examples are closer together and dissimilar ones are farther apart.

**Result:**

| Dataset | Result |
|---|---|
| CUHK | 64.87 |
| Person face Sketch(Kaggle) | 72.84 |

Table 3.3.1 Accuracy Result on Triplet Network

# CHAPTER 4

# CONCLUSION

The limited training data are currently the primary obstacle to adopting deep learning in face sketch recognition.

To map joint features in the first method, we paired a sparse encoder-decoder network with a Siamese network. Because artists have different painting techniques, most convolutional networks find it difficult to be trained because there aren't enough images. For training, we employed the Siamese network structure, which can pick up a lot of information from a short amount of input. This is a result of the Siamese network's utilization of image pairs, which are made up of a facial sketch and a photo. We created a large number of input images, either sketches or pictures, of faces in order to train every image pair for recognition. By doing so, overfitting was prevented and the quantity of training datasets was increased.

A unique triplet model was our second model. Low recognition accuracy was the result of short datasets and poor loss function convergence, even though deep learning techniques were used to extract many features. Because drawings are reduced copies of images, features retrieved from photos and their associated sketches in Siamese networks frequently don't match well. In order to solve this, we created an attention model that concentrates on identifying features from the same regions in the sketch and the image. In this approach, the differences between the sketch and the photo are reduced when they are combined into a similar feature space. A channel block and two distinct spatial blocks—one for processing pictures and the other for processing sketches—made up the attention model.

## REFERENCES

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2282.

[2] Klum, Scott, Hu Han, Anil K. Jain, and Brendan Klare. "Sketch based face recognition: Forensic vs. composite sketches." In *2013 international conference on biometrics (ICB)*, pp. 1-8. IEEE, 2013.

[3] Barbadekar, Ashwini, and Prajakta Kulkarni. "A survey of face recognition from sketches." *International Journal of Latest Trends in Engineering and Technology (IJLTET)* 6, no. 3 (2016): 150-158.

[4] Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. "Matching networks for one shot learning." *Advances in neural information processing systems* 29 (2016).

[5] Cao, Bing, Nannan Wang, Xinbo Gao, Jie Li, and Zhifeng Li. "Multi-margin based decorrelation learning for heterogeneous face recognition." *arXiv preprint arXiv:2005.11945* (2020).

[6] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[7] Martinez, Aleix M., and Avinash C. Kak. "Pca versus lda." *IEEE transactions on pattern analysis and machine intelligence* 23, no. 2 (2001): 228-233.

[8] Tang, Xiaoou, and Xiaogang Wang. "Face sketch recognition." *IEEE Transactions on Circuits and Systems for video Technology* 14, no. 1 (2004): 50-57.

[9] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2, no. 1-3 (1987): 37-52.

[10] Chelali, Fatma Zohra, A. Djeradi, and R. Djeradi. "Linear discriminant analysis for face recognition." In *2009 International Conference on Multimedia Computing and Systems*, pp. 1-10. IEEE, 2009.

[11] Sharma, A., Jacobs, D.W., 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On. IEEE, pp. 593–600.

[12] Kan, Meina, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. "Multi-view discriminant analysis." *IEEE transactions on pattern analysis and machine intelligence* 38, no. 1 (2015): 188-194.

[13] Ahonen, Timo, Abdenour Hadid, and Matti Pietikäinen. "Face recognition with local binary patterns." In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pp. 469-481. Springer Berlin Heidelberg, 2004.

[14] Lindeberg, Tony. "Scale invariant feature transform." (2012): 10491.

[15] Huang, Gao, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. "Deep networks with stochastic depth." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646-661. Springer International Publishing, 2016.

[16] Li, Dangwei, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. "Learning deep context-aware features over body and latent parts for person re-identification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 384-393. 2017.

[17] Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "Deepface: Closing the gap to human-level performance in face verification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701-1708. 2014.

[18] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. "Deep learning face representation from predicting 10,000 classes." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1891-1898. 2014.

*[19]* Güçlütürk, Yağmur, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition." *In Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016".*

[20] Liu, QingShan, Rui Huang, HanQing Lu, and SongDe Ma. "Kernel- based nonlinear discriminant analysis for face recognition." *Journal of Computer Science and Technology* 18, no. 6 (2003): 788-795.

[21] Parkhi, Omkar, Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[22] Bromley, Jane, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. "Signature verification using a" siamese" time delay neural network." *Advances in neural information processing systems* 6 (1993).

[23] Klare, Brendan, and Anil K. Jain. "On a taxonomy of facial features." In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1-8. IEEE, 2010.

PAPER NAME

Face_sketch_recognition.pdf

WORD COUNT

7337 Words

CHARACTER COUNT

40522 Characters

PAGE COUNT

33 Pages

FILE SIZE

802.6KB

SUBMISSION DATE

May 30, 2024 11:46 AM GMT+5:30

REPORT DATE

May 30, 2024 11:47 AM GMT+5:30

Varsha

● 15% Overall Similarity

**The combined total of all matches, including overlapping sources, for each database.**

- **10% Internet database**
- **5% Publications database**
- **Crossref database**
- **Crossref Posted Content**
- **database9% Submitted Works database**

● Excluded from Similarity Report

- **Bibliographic material**
- **Small Matches (Less then 8 words)**

ANNEXURE-IV

# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

## PLAGIARISM VERIFICATION

Title of the Thesis FACE SKETCH RECOGNITION USING DEEP LEARNING Total Pages 33 Name of the Scholar OM AMRIT Supervisor (s)

(1) Dr. Varsha Sisaudia
(2)____________________________________________
(3)____________________________________________

Department INFORMATION TECHNOLOGY, DELHI TECHNOLOGICAL UNIVERSITY

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used:______Turnitin________ Similarity Index:__15 %___, Total Word Count: 7337

Date: 31/05/2024

Varsha

**Candidate's Signature** **Signature of Supervisor(s)**